COGNITIVE SCIENCE SOCIETY July 25-26, 2005, Stresa, Italy, pp. 45-49

TOWARD SOCIAL MECHANISMS OF ANDROID SCIENCE
A COGSCI 2005 WORKSHOP

Synthesizing Affect with an Analog Vocal Tract: Glottal Source

Michael C. Brady (mcbrady@indiana.edu)
Depts. of Linguistics and Cognitive Science
322 Memorial Hall, Indiana University
Bloomington, IN 47405 USA

Abstract

An analog speech synthesizer is constructed based on the source-filter model of the human vocal tract. The artificial tract was developed in part for studies on emotional and paralinguistic content in speech synthesis and for eventual studies on speech interaction. In this study, the glottal open quotient (OQ) of the waveforms generated as the source (weak to strong high frequency spectrum), the fundamental frequency (F0), and a jitter parameter were varied as the filter shape or tongue posture was randomly altered. Subjects were first asked to rate caricature facial expressions on affective attributes and then to rate goodness of fit between the faces and the sound stimuli. Association scores between attributes and sounds based on these ratings were calculated. Results indicate that in the context of isolated sound segments, OQ is the strongest determinant of perceived android arousal and F0 is the strongest determinant of perceived android valence. Further results point to dependencies between the three varied parameters. Findings are consistent with related research and help to establish a foundation for future work with the artificial tract. A concluding discussion remarks on how this study and the project itself speaks to other research on speech and to android science in general. The rich visual cues and the more ecologically realistic acoustic radiation characteristics of an analog synthesizer are strategic to a science where naturalistic mechanisms for social interaction are needed.

Introduction

When a dog wags its tail, raises it, or curls it between its legs, it communicates with us. We infer the dog's affective or emotive state and we can thereby relate. People transmit similar affective cues. Facial expressions and displayed interpretations of one another's internal states during dialog are integral to our interpersonal relations. Our empathy-based expectations are also manifest in the human-machine interaction. An android that may "wag or hang its acoustic tail" will be able to relate with people more naturally and more effectively. Such an android will allow for controlled studies of affect mirroring, social referencing, and other forms of empathetic behavior in aural communication.

A great deal of work has been conducted on paralinguistic content in speech. Sarcasm, humor, confidence and truth, dialect, gender, dominance and submission, age, self-concept, prosody and conversational turn taking, and other topics are important issues when considering the android speech interface. This study focuses on establishing ways in which the percepts of valence and arousal may be achieved in voice synthesis based on manipulations of the glottal source. How these primary affective dimensions can be realized when producing speech sounds is fundamental to some of the more complex topics in speech interaction.

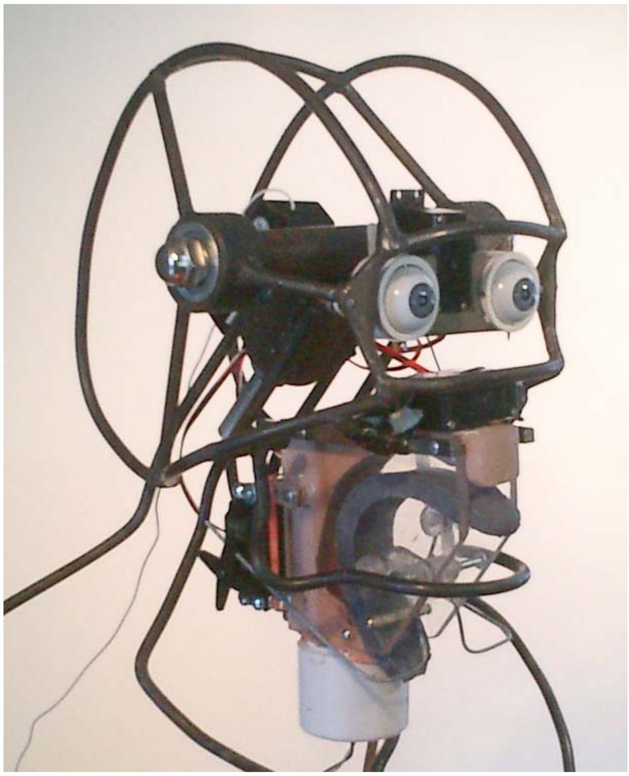

Figure 1: Android head with analog vocal tract.

Previous Work

Emotion and affect in speech is a vast topic with many unresolved issues. A main problem is that the context of a social situation largely determines how acoustic cues are to be interpreted. Additionally, modern research is careful to differentiate emotion from mood, attitude, and personality. The very language we use to talk about affect can be imprecise and questions arise about what emotion even is with respect to what we are working to model. It has become increasingly difficult and perhaps even naive to discuss emotion in speech in terms of isolated and idealized cues. For further perspective on such topics, see (Campbell, 2004). In the face of these complexities, there are some fundamentals we can consider.

It is clear that listeners are able to distinguish some specific emotions in the voice, such as anger, grief and elation. Even children as young as three or four years of age are able to identify anger, happiness and sadness from intonation patterns and anger and happiness from vocal cues alone (Hortacsu and Ekinci, 1992). Moreover, emotions can be recognized in speech segments as short as 60 milliseconds (Bachorowski and Owren, 1995). The basic universal emotions have been the target of much work on emotional content in speech over a few underlying acoustic parameters. For the expressions of happy, sad, angry, scared and disgusted, researchers have typically looked to the mean, range, and variability of fundamental frequency and intensity, at speech rate and at voice quality (noise and portion of high spectral content to low spectral content) as cues to inform perception. The following is a loose summary of acoustic traits in speech that tend to correspond to perceived affective states (Murray and Arnott, 1993; Pittam, 1994; Scherer, 1986; Schröder, 2003):

- Anger is typically characterized by an increase in F0 and mean intensity as well as by an increase in high frequency energy.
- Fear is also characterized by an increase in F0 and high frequency energy while mean intensity does not seem to be as much of a factor.
- Sadness typically involves a decrease in F0 and a decrease in intensity as well as a decrease in high frequency energy.
- Joy, like anger, is typified by increases in F0 and increases in intensity.
- Disgust cues in voice are not well agreed upon and seem to depend on the circumstances of the study.

Perhaps it is best to consider these generalizations in the context of physiology. Air passes from the lungs through the vocal cords and out the mouth. A person's affective state (e.g., fight-or-flight activation of peripheral nervous system) can influence how much oxygen is metabolized and this influences breathing rate. Changes in breath rate will influence the speed and character with which an utterance is produced. At the segmental level, quickened respiration leads to increases in subglottal air pressure which, in turn, puts more strain on the vocal cords. This impacts how the cords vibrate. In response, muscular tension on the cords tends to spontaneously increase and this raises F0. What's more, consider that the extrinsic muscles of the larynx, (muscles responsible for the height of the larynx), are innervated by a branch of the cranial nerve (VII) that innervates the muscles responsible for facial expressions. This same nerve also provides the parasympathetic connections that control a local saliva gland's secretions. Vocal fold lubrication impacts vocal fold response, larynx height impacts the filter characteristic of the vocal tract, and everything basically influences everything else. For continued discussion on the incredibly complex dynamics of vocalization, see (Bless and Abbs, 1983; Laver 1980). The important point to realize is that we as listeners seem to be able to identify the physiological state of a speaker's vocal system based on acoustic cues. Like facial expressions, these cues surely enter into our shared communication systems and may be employed as expressive markers, even in the absence of strong emotion.

Speech Synthesis

In 1876, Charles Wheatstone demonstrated a famous analog speaking machine (based on a 1791 invention by Wolfgang von Kempelen). It produced vowels and most of the consonants. By 1922, speech synthesis by electrical circuitry was invented and by 1939, Homer Dudley demonstrated his "Voder," a speech synthesizer based on a bank of electrical bandpass filters.

Virtually all of today's work in speech synthesis is performed on the digital computer where electrical output is sent to a loudspeaker. The three main approaches to digital synthesis are concatenative synthesis, formant synthesis, and articulatory synthesis. Concatenative synthesis involves chopping pre-recorded real speech into segments and splicing those segments together as necessary. Formant and articulatory synthesis are based on generative algorithms. Formant synthesis (like the Voder) produces speech sounds by passing a source signal through a series of electric or digital filters where the source and filter values change 20 to 30 times per second. Articulatory synthesis builds on this. Here, parameters representing positions of articulators (such as tongue, lip, and jaw) are varied and resulting simulated vocal tract shapes determine the source and filter values of a formant synthesizer. Because of its naturalistic origins, concatenative synthesis has enjoyed most of the glory in producing today's best sounding artificial speech. Problems in creating realistic intonation patterns with the earlier strategies of concatenative synthesis, called 'diaphone synthesis', have been addressed by 'unit-selection' techniques and more grand-scale concatenative approaches, as well as by the development of concatenative-formant hybrid types of models.

With the advent of sophisticated motors, microcontrollers, and robotics technologies, we might expect to see a renewed push in the analog direction. For instance, one approach (Sawada & Nakamura, 2004) uses compressed air to vibrate artificial vocal cords. The resulting buzz sound is then shaped as it passes through a motor-actuated silicone tube. An analog device reduces the computational overhead of articulatory synthesis because it bypasses the complexity of mapping dynamically related articulator positions to formant synthesis values. Furthermore, more ecologically realistic speech may be produced by acoustic mechanisms.

Experiment

An analog speech synthesizer was constructed based on the source-filter model of the human vocal tract. The filter is realized as a resonant chamber built around a flexible tongue-like apparatus and the source is realized as a simple electric speaker attached to seal one end of the chamber. Two servo motors actuate the tongue via two independent lever mechanisms. This allows the tongue to move with two degrees of freedom (front to back, high to low), and to mimic the way the human tongue moves when producing vowel sounds. Example movies of the head and tract are available at (http://mypage.iu.edu/~mcbrady/tract).

Stimuli

Three parameters were varied in generating buzz sounds from the electric speaker: fundamental frequency (F0), glottal open quotient (OQ), and a jitter-shimmer parameter (JS). The posture of the tongue was continuously and randomly altered between three vowel shapes (/a/, /i/, /u/) during the course of the experiment. This was done in anticipation that results should be applicable to future research where the tongue movement is the variable, as well as to aid subjects in hearing the stimuli as speech sounds. Each sound segment or vocalization lasted 500 ms. A speech-like standardized amplitude envelope was applied to the waveforms so that onsets and offsets of stimuli sounded more natural. There were two F0 conditions, three OQ conditions, and two JS conditions, as described below. In total, there were 2 x 3 x 2 = 12 sound stimuli.

F0

F0 is the frequency at which the vocal folds vibrate during speech. The artificial tract is modeled as that of a typical male (length of 17 cm from end to end). The characteristic F0 for a male is near 120 Hz. 140 Hz and 100 Hz mark a reasonable male F0 range and these were the values used for the F0 high and low conditions, respectively.

OQ

Glottal open quotient refers to the percentage of time within the glottal cycle that the vocal folds remain open. Vocal folds vibrate periodically in a way dependent on the Bernoulli effect and on vocal cord tension. When air pressure is greater below the glottis, the vocal folds close more rapidly and remain closed for a greater portion of their vibratory cycle. A Fourier analysis of this higher pressure waveform indicates greater high frequency energy. When there is less air pressure below the glottis, the vocal folds remain open for the greater portion of their vibratory cycle and they come back together more slowly, approximating something akin to a sinusoidal function. A Fourier analysis of this lower pressure waveform reveals reduced high frequency components. A short OQ of 0.1, a neutral OQ of 0.5, and a long OQ of 0.8 were used in this study. These OQs are termed “creaky,” “modal,” and “breathy,” respectively. Fig. 2 depicts these waveforms. ‘Creaky’ resonates in the tract with more ‘brightness’ and with stronger high frequency components while ‘breathy’ resonates more sinusoidally. The analogy of a saxophone helps to conceptualize this; if one blows hard on the mouthpiece, the resulting timbre is brilliant and even shrill or harsh, and when one blows softly the resulting sound can be described as ‘breathy’, ‘smooth’, or ‘mellow’.

JS

Natural glottal pulses are not perfectly periodic, they exhibit period-to-period variation in both duration and amplitude. Jitter refers to duration variation while shimmer refers to amplitude variation. These variations were combined into a single parameter, JS, where duration variation is inversely related to amplitude variation. Specifically, a random number, r, is generated for each glottal pulse within a range, -K to K. The duration of that pulse is multiplied by 1+r (pulse is truncated or zero-padded) and the amplitude of each sample in the pulse is multiplied by 1-r. Two JS conditions were used in this study, a normal condition (K = .01) and a high condition (K = .05). In the normal condition, it is difficult to detect JS. In the high condition, JS is quite readily identified as being present.

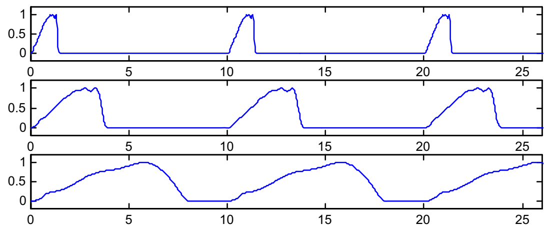

Figure 2: Creaky (top), modal (middle), and breathy (bottom) vocal source waveforms

Response Prompts

Fig. 3 presents eight caricatures of facial expression used for subject response ratings (Faigin, 1990). Male faces were used because the artificial tract is based on a typical male vocal tract length. The faces were selected to capture ranges of emotion and arousal appropriate for the modeling of android affect. Drawings rather than photographs were chosen so that subjects might more readily appreciate the abstract nature of the task.

Method

Twenty college age male and female subjects were recruited in Bloomington, Indiana. All had normal hearing and vision and were native speakers of English.

Subjects were first asked to rate the facial expressions across seven attributes (*happy, sad, angry, scared, disgusted, sleepy/passive, roused/activated*) on an eleven-point scale (0: no trace of the attribute perceived, 10: the attribute was perceived very much). For each trial, a subject saw a face on the computer screen. Below the face was the question: “how __ does this face look?” where the attribute word from the same list as above filled in the blank. Subjects clicked on a number from 0 to 10 and then clicked *‘enter’*. There were 7 x 8 = 56 trials presented in pseudo-random order for this block. Subjects were then asked to rate the 12 speech sounds in relation to the facial expressions on the same eleven-point scale. For each trial a subject was presented with a face, below which was the question “how well does this face match the vocal sound?” (0: not at all, 10: very much). After a brief pause the stimulus sound would be synthesized on the android vocal tract and the subject could then either enter a rating or click to hear the sound again. There were 12 x 8 = 96 trials presented in pseudo-random order for this second and final

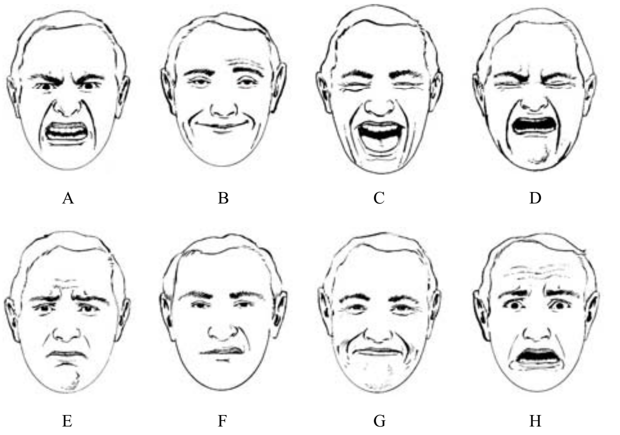

Figure 3: Eight facial expressions as response prompts. Taken from “The Artist’s Complete Guide to Facial Expression” by Gary Faigin.

block. At the beginning of both blocks there were 5 throw away trials to familiarize the subject with the equipment and task. The experiment lasted about fifteen minutes.

Results

Table 1 presents the mean attribute ratings for the faces from the first half of the experiment. Overall response provides a basis by which to discuss subject ratings of sound stimuli. Note that faces A, C, D, and H are generally judged to be highly aroused or active, while faces B, E, F, and G are judged as less aroused or passive. Faces B, C, and G are judged as happy while faces D and E are more commonly judged to be sad.

Fig. 4 presents the results of the speech stimuli ratings averaged over subjects for each of the three varied sound synthesis dimensions. Results for each dimension are pooled over the other two respective dimensions. For example, all stimuli with an F0 of 140 Hz, regardless of JS or OQ were rated on average 4.4 in response to face A.

In the top graph of Fig. 4, faces A, D, E, and F (negative valence faces) are rated as more representative of the speech sounds that have lower F0s. Faces B, C, and G (the happy or positive valence faces) are rated as more representative of the speech sounds with higher F0s. The middle graph of Fig. 4 indicates a slight effect of JS where faces A, D, E, and H receive higher ratings in response to sounds with high JS. In the bottom graph of Fig. 4, faces A, C, D, and H (high arousal faces) receive higher match ratings as OQ is decreased while faces B, E, F, and G (lower arousal faces) receive higher match ratings when OQ is increased.

Table 1: Attribute ratings pooled over subjects.

		Happy	Sad	Angry	Scared	Disgusted	Passive	Activated
Face	A	0.2	2.3	9.2	2.3	6.3	0.5	8.5
	B	7.1	0.6	0.5	0.5	0.5	7.3	2.5
	C	8.9	0.4	0.8	1.7	1.1	0.9	9.0
	D	0.2	7.3	5.5	3.4	5.5	2.5	7.7
	E	0.4	7.5	2.8	3.7	3.1	6.5	3.9
	F	1.5	2.2	4.5	1.1	5.7	5.7	2.8
	G	7.7	0.7	0.6	0.7	0.7	4.5	4.7
	H	0.5	3.5	3.5	8.1	4.7	1.2	8.5

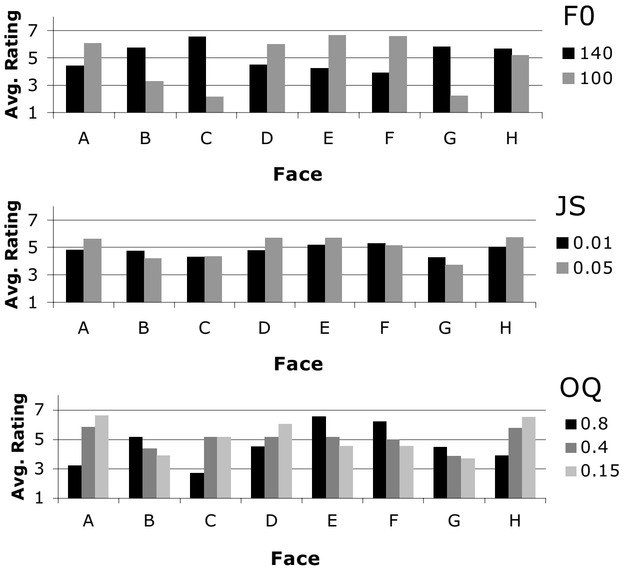

Figure 4: Ratings pooled over glottal source dimensions.

Fig. 5 provides an alternative visualization of the data. Each of the 12 sound stimuli is plotted in valence-versus-arousal space based on word-to-sound association scores. An association score for a word given a sound stimulus, $A_{W|S}$, is taken as the product of the mean word-to-face rating, w_f, for a face (from first phase of experiment) and that face's mean sound-to-face rating, s_f, (from second phase of experiment), averaged over faces.

$$A_{W|S} = \frac{\sum_{f=1}^{N} w_f s_f}{N} \tag{1}$$

The top graph of Fig. 5 plots association scores of stimuli for happy against active while the bottom graph plots sad against passive. Triangles indicate stimuli with higher F0s (versus circles) while grayscale saturation indicates extent of OQ. Stimuli marked with '^' are of the high JS condition.

Overall, people seem to be more responsive to F0 as a valence cue and to OQ as an arousal cue. This is apparent upon analysis of both Fig. 4 and Fig. 5, where speech sounds with an F0 of 140 Hz were responded to as happier or less sad (and vice versa for the 100 Hz sounds), and sounds with a large OQ were perceived as less activated or more passive (and vice versa for small OQ sounds). Jitter

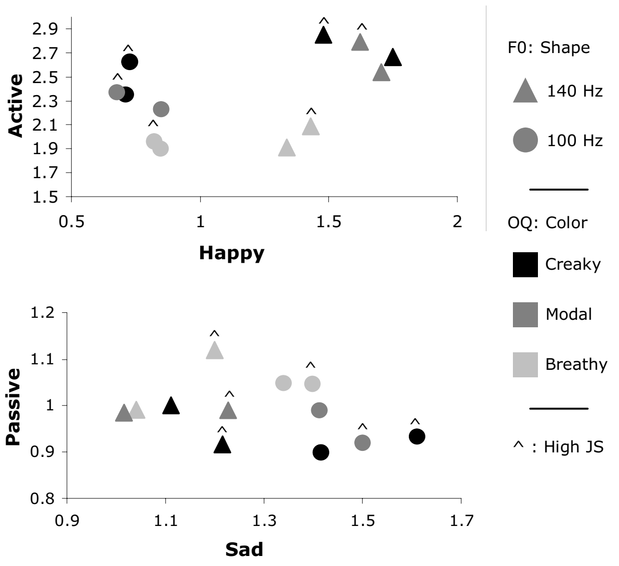

Figure 5: Stimuli plotted in valence versus arousal space.

had a minor effect where the high JS sounds were generally heard as less happy and more aroused. One interpretation may be that JS indicates stressful or worried speech. Lastly, there was an overall bias for subjects to rate the stimuli as representing the positive valence faces less well. This is observed in Fig. 4 where faces A, D, E, F, and H (not happy) received generally higher fit ratings on the whole than faces B, C, and G (happier).

Discussion

Results from this study form part of a much larger picture. In essence, the analog tract is an articulatory synthesizer. Understanding the effects of individual parameters is a first step in developing dynamic control algorithms. Emotional and affective voice has not been a relative concern to researchers in articulatory synthesis. Perhaps this is because their systems are still not very good at unsupervised production of intelligible words, and success seems to be measured in terms of this. When considering speech production from an android science perspective however, a rudimentary question must be asked: how can we expect to build articulatory devices to effectively speak words and strings of words when we hardly know how to build such devices to make expressive grunts and intonation patterns?

The Chomskian view in mainstream linguistics (lightning summary: extract symbols from the speech stream, do computations on those symbols, discard the rest) has distanced the study of language from its roots in vocal communication (Cowley & MacDorman, 1995; Port, 2005). However, mainstream theories of language will eventually have to address the issues of embodiment and interactivity. Social influences are shown to impact speech production, see Suzuki & Katagiri's (2005) study of prosodic adaptation, also at this workshop. Also, context has an undeniable impact on meaning. Ultimately speech and language, and more assertively, language in mind, must rely on the situations and relationships of the speakers and listeners involved.

The analog vocal tract of this study is intended to open a dialog on the need for an android paradigm in linguistics. Emotion, social cues, ecological factors, and more will need to be incorporated into theories of cognitive representation for speech. Situating the tract in an anthropomorphic head allows the tract to be further taken as a social mechanism, facilitating the development of interactive speech synthesis techniques. Eventually these vocal abstractions may offer fundamental new insights into theories of speech and cognitive function.

Acknowledgements

Thanks go to Robert Port, Diane Kewley-Port, and to John Paolillo for their valuable comments on this work. Thanks also go to Galo Moncayo and Steve Lauderback of the Indiana University School of Fine Arts for aid in fabricating the android head.

References

Bachorowski, J., & Owren, M. (1995). Vocal expression of emotion. *Psychological Science, 6,* 219-224.

Bless, D. M., & Abbs, J. H. (1983). *Vocal fold physiology: contempory research and clinical issues,* San Diego: College Hill Press.

Campbell, N. (2004). Specifying affect and emotion for expressive speech synthesis. *Lecture Notes in Computer Sciences, 2945,* 395-406.

Cowley, S., & MacDorman, K. (1995). Simulating conversations: The communion game. *AI & Society, 9*(3), 116-137.

Faigin, G. (1990). *The complete artist's guide to facial expresion*. New York: Watson-Guptill Publications.

Hortacsu, N., & Ekinci, B. (1992). Children's reliance on situational and vocal expression of emotions. *Journal of Nonverbal Behavior, 16,* 231-247.

Laver, J. (1980), *The Phonetic Description of Voice Quality,* Cambridge: Cambridge University Press

Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech. *Journal of Acoustical Society of America, 93*(2), pp. 1097-1108.

Pittam, J. (1994). *Voice in social interaction*. Thousand Oaks: Sage.

Port, R. (2005). *Social, sensory and symbolic aspects of phonology* (in prep.) http://www.cs.indiana.edu/~port

Sawada, H., & Nakamura, M. (2004). Mechanical voice system and its singing performance. *IEEE/RSJ Intl. Conf. On Intelligent Robots ans Systems*, pp. 1920-1925.

Scherer, K. (1986). Vocal affect expression. *Psychological Bulletin, 99(2),* 143-165.

Schröder, M. (2003). Experimental study of affect bursts, *Speech Communication, 40*(1-2). pp 99-116.

Suzuki, N. & Katagiri, Y. (2005). Prosodic adaptation in human-computer interaction. *Proceedings of the 2005 Annual Cognitive Society Conference Workshop: Toward Social Mechanisms of Android Science.*